

ACTION RECOGNITION METHODS AND APPARATUSES, ELECTRONIC DEVICES, AND STORAGE MEDIA

[0001] This patent application claims priority to Chinese Patent Application No.
5 201910252534.6 entitled “action recognition methods and apparatuses, electronic
devices, and storage media” and filed on March 29, 2019, the entire contents of which
are incorporated herein by reference.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of computer vision technologies,
10 and in particular, to action recognition methods and apparatuses, electronic devices, and
storage media.

BACKGROUND

[0003] In the field of computer vision, the action recognition is always an issue of
great interest. In general, researches on action recognition focus on a temporal
15 sequential feature of a video, and some actions that may be recognized in accordance
with body key points.

SUMMARY

[0004] The embodiment of the present disclosure provides an action recognition
technology.
20 [0005] According to an aspect of the embodiments of the present disclosure, an action
recognition method is provided, including: obtaining mouth key points of a face based
on a face image; determining, based on the mouth key points, an image in a first region
involves at least part of the mouth key points and comprises an image of an object

interacting with mouth; determining whether a person corresponding to the face image is smoking based on the image in the first region.

[0006] According to another aspect of the embodiments of the present disclosure, an action recognition device is provided, including: a mouth key point module, configured to obtain mouth key points of a face based on a face image; a first region determining module, configured to determine, based on the mouth key points, an image in a first region involves at least part of the mouth key points and comprises an image of an object interacting with mouth; a smoking recognizing module, configured to determine whether a person corresponding to the face image is smoking based on the image in the first region.

[0007] According to another aspect of the embodiments of the present disclosure, an electronic device is provided, including a processor, wherein the processor includes an action recognition apparatus according to any of the foregoing embodiments.

[0008] According to still another aspect of the embodiments of the present disclosure, an electronic device is provided, including: a memory for storing executable instructions; and a processor, configured to communicate with the memory to execute the executable instructions to perform operations in the action recognition method in any of the above embodiments.

[0009] According to another aspect of the embodiments of the present disclosure, a computer readable storage medium is provided for storing computer readable instructions, wherein the instructions are executed to perform the operations in the action recognition method according to any of the above embodiments.

[0010] According to another aspect of the embodiments of the present disclosure, a computer program product is provided, which includes computer readable codes, wherein the computer readable codes are running on a device to cause a processor in

the device to execute instructions for implementing the action recognition method according to any of the above embodiments.

[0011] Based on the action recognition methods and apparatuses, electronic devices, and storage media according to the above embodiments of the disclosure, mouth key points of a face is obtained based on a face image; an image in a first region is determined based on the mouth key points, and the image in the first region includes at least part of mouth key points and an image of an object interacting with mouth; and based on the image in the first region, it may determine whether the person corresponding to the face image is smoking. By recognizing the image in the first region determined with the mouth key points and accordingly determining whether the person corresponding to the face image is smoking, the recognition range is effectively reduced and attention may be focused on the mouth and the object interacting with mouth, thereby improving the detection rate, reducing the false detection rate, and improving the accuracy of smoking recognition.

[0012] The technical solution of the present disclosure will be further described in detail below through the drawings and embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The accompanying drawings, which form a portion of the description, describe embodiments of the present disclosure, and together with the description serve to explain the principles of the present disclosure.

[0014] The present disclosure will be more clearly understood from the following detailed description with reference to the accompanying drawings.

[0015] FIG. 1 is a schematic flowchart of an action recognition method according to an embodiment of the present disclosure.

[0016] FIG. 2 is another schematic flowchart of an action recognition method according to an embodiment of the present disclosure.

[0017] FIG. 3A is a schematic diagram of a first key point obtained via recognition in an action recognition method according to an embodiment of the present disclosure.

5 [0018] FIG. 3B is a schematic diagram of a first key point obtained via recognition in an action recognition method according to another embodiment of the present disclosure.

[0019] FIG. 4 is another schematic flowchart of an action recognition method according to an embodiment of the present disclosure.

10 [0020] FIG. 5 is a schematic diagram illustrating an alignment operation performed on an object interacting with mouth in an action recognition method according to another embodiment of the present disclosure.

[0021] FIG. 6A illustrates a captured original image in an action recognition method according to an embodiment of the present disclosure.

15 [0022] FIG. 6B is a schematic diagram of detecting a face frame in an action recognition method according to an embodiment of the present disclosure.

[0023] FIG. 6C is a schematic diagram of a first region determined based on key points in an action recognition method according to an embodiment of the present disclosure.

20 [0024] FIG. 7 is a schematic structural diagram of the action recognition apparatus according to an embodiment of the present disclosure.

[0025] FIG. 8 is a schematic structural diagram of an electronic device applicable to a terminal device or a server according to an embodiment of the present disclosure.

DETAILED DESCRIPTION OF THE EMBODIMENTS

25 [0026] Various exemplary embodiments of the present disclosure will now be

described in detail with reference to the accompanying drawings. It should be noted that the relative arrangements, numerical expressions, and numerical values of the components and steps set forth in these embodiments do not limit the scope of the present disclosure unless specifically stated otherwise.

5 **[0027]** For convenience of description, the dimensions of the various portions shown in the figures are not drawn according to actual proportional relationships.

[0028] The following description of at least one exemplary embodiment is practically merely illustrative, in no way limiting of the present disclosure and its application or use.

10 **[0029]** Techniques, methods, and apparatuses known to those of ordinary skill in the relevant art may not be discussed in detail, but the techniques, methods, and apparatuses should be considered as portion of the description unless otherwise specified.

[0030] It should be noted that like reference signs and letters denote like items in the following drawings, and therefore, once a certain item is defined in one figure, no
15 further discussion thereof is needed in the following drawings.

[0031] Embodiments of the present disclosure may be applicable to computer systems/servers, which may operate with numerous other general purpose or special purpose computing system environments or configurations. Examples of well-known computing systems, environments and/or configurations suitable for use with computer
20 systems/servers include, but are not limited to, personal computer systems, server computer systems, thin clients, thick clients, hand-held or laptop devices, microprocessor-based systems, set-top boxes, programmable consumer electronics, network personal computers, minicomputer systems, mainframe computer systems, and distributed cloud computing technology environments including any of the above, and
25 the like.

[0032] The computer system/server may be described in the general context of computer system-executable instructions, such as program modules, executed by the computer system. In general, program modules may include routines, programs, target programs, components, logic, data structures, etc. that perform particular tasks or

5 implement particular abstract data types. The computer system/server may be implemented in a distributed cloud computing environment in which tasks are performed by a remote processing device linked via a communication network. In a distributed cloud computing environment, the program modules may be located on a local or remote computing system storage medium including a storage device.

10 [0033] FIG. 1 is a schematic flowchart of an action recognition method according to an embodiment of the present disclosure. The method may be applicable to an electronic device, as shown in FIG. 1, the method of this embodiment includes the following steps.

[0034] At step 110, mouth key points of a face are obtained based on a face image.

[0035] The mouth key points in the embodiment of the present disclosure may be

15 performed by labelling the mouth on the face. The mouth key points may be obtained by any implementable face key point recognition method in the prior art. For example, some face key points on the face are recognized by using a deep neural network, the mouth key points then are obtained by separating from the face key points. For another example, the mouth key points are recognized directly by using deep neural network.

20 The embodiment of the present disclosure does not limit the specific manner of obtaining a mouth key point.

[0036] In an example, the step 110 may be performed by the processor invoking corresponding instructions stored in the memory, or by the mouth key point module 71 executed by the processor.

25 [0037] At step 120, an image in a first region is determined based on the mouth key

points.

[0038] In an example, the image in the first region involves at least part of the mouth key points and comprises an image of an object interacting with the mouth. The action recognition in the embodiment of the present disclosure is mainly used for recognizing whether a person corresponding to an image is smoking, and since the smoking action is realized by contacting the mouth with the cigarette, not only part (a portion) or all of the mouth key points are included in the first region, but also an object interacting with mouth may be included. When the object interacting with mouth is the cigarette, it is determined that the person corresponding to the image may smoke. As an example, the first region in the embodiment of the present disclosure may be of any shape such as a rectangle or a circle determined based on the center of the mouth. In the embodiment of the present disclosure, the shape and size of the image in the first region are not limited, taking an interaction such as cigarette and rod sugar that may be in contact with the mouth in the first region as a standard.

[0039] In an example, step 120 may be performed by the processor invoking the corresponding instructions stored in the memory, or by the first region determining module 72 executed by the processor.

[0040] At step 130, it is determined whether a person corresponding to the face image is smoking based on the image in the first region.

[0041] In the embodiment of the present disclosure, it is determined whether the person corresponding to the face image is smoking by recognizing whether the object interacting with mouth included in a region in the vicinity of the mouth is cigarette, centralizes a point of interest in the vicinity of the mouth, which reduces the probability that other irrelevant images interfere with the recognition result, and improves the accuracy of recognizing smoking action.

[0042] In an example, the step 130 may be performed by the processor invoking the corresponding instructions stored in the memory, or by the smoking recognizing module 73 executed by the processor.

[0043] Based on the action recognition method according to the above embodiments of the present disclosure, the mouth key points of the face is obtained based on the face image; the image in the first region is determined based on the mouth key points, and the image in the first region includes at least part of mouth key points and an object interacting with mouth. Based on the image in the first region, it is determined whether the person corresponding to the face image is smoking, and the image in the first region determined by the mouth key points is identified to determine whether the person corresponding to the face image is smoking, which reduces the recognition range, concentrates attention on the mouth and the object interacting with mouth, thereby improving the detection rate, reducing the false detection rate, and improving the accuracy of smoking recognition.

[0044] FIG. 2 is another schematic flowchart of an action recognition method according to an embodiment of the present disclosure. As shown in FIG. 1, the method in this embodiment includes the following steps.

[0045] At step 210, mouth key points of a face are obtained based on a face image.

[0046] At step 220, an image in a first region is determined based on the mouth key points.

[0047] At step 230, at least two first key points of the object interacting with mouth are obtained based on the image in the first region.

[0048] In an example, key point extraction on the image in the first region may be performed via a neural network, so as to obtain at least two first key points of the object interacting with mouth, and the first key points may appear as a straight line in the first

region (for example, extracting cigarette key points from an axis in the cigarette) or two straight lines (for example, extracting cigarette key points from two sides of the cigarette), etc.

5 **[0049]** At step 240, the image in the first region is screened based on the at least two first key points.

[0050] The screening operation is to select out the image in the first region in which the object interacting with mouth and having a length greater than or equal to a preset value is involved.

10 **[0051]** In an example, the length of the object interacting with mouth in the first region may be determined by at least two first key points on the obtained object that interacts with the mouth. When the length of the object interacting with mouth is small (for example, the length of the object interacting with mouth is less than a preset value), the object interacting with mouth included in the first region is not necessarily a cigarette. In this case, it may be determined that the image in the first region does not include a
15 cigarette. Only when the length of the object interacting with mouth is large (for example, the length of the object interacting with mouth is greater than or equal to a preset value), it is determined that a cigarette may be included in the image in the first region.

20 **[0052]** At step 250, in response to that the image in the first region passes the screening, it is determined whether the person corresponding to the face image is smoking based on the image in the first region.

[0053] In the embodiment of the present disclosure, the above screening operation determines an image in part of the first region, and the image in the portion of the first region includes an object interacting with mouth and has a length reaching a preset
25 value. Only when the length of the object interacting with mouth reaches a preset value,

It is determined that the object interacting with mouth may be a cigarette. In this step, it is determined whether the person in the face image is smoking with respect to the image in the first region which passes the screening. In other words, when an object interacting with mouth has a length greater than the preset value, it is determined
5 whether the object interacting with mouth is a cigarette; to determine whether the face in the face image is smoking.

[0054] In an example, step 240 includes:

[0055] key point coordinates corresponding to at least two first key points in the image in the first region are determined based on the at least two first key points;

10 **[0056]** the image in the first region is screened based on the key point coordinates corresponding to the at least two first key points.

[0057] After the at least two first key points of the object interacting with mouth are obtained, it is not possible to completely determine whether the person in the face image is smoking, maybe other similar objects (e.g., rod sugar or other elongated objects, etc.)
15 are contained in the mouth. The cigarette generally has a certain length to determine whether the first region includes the cigarette. In the embodiment of the present disclosure, the key point coordinates of the first key point are determined, and the length of the object interacting with mouth in the first region image is determined based on the key point coordinates of the first key point in the first region, thereby determining
20 whether the person in the face image is smoking.

[0058] In an example, the image in the first region is screened based on the key point coordinates corresponding to the at least two first key points includes:

[0059] a length of the object interacting with mouth in the image in the first region is determined based on the key point coordinates;

25 **[0060]** in response to the length of the object interacting with mouth being greater

than or equal to the preset value, it is determined that the image in the first region passes the screening.

[0061] In an example, in order to determine the length of the object interacting with mouth, after the key point coordinates of the at least two first key points are obtained, the at least two first key points at least include a key point on an end of the object near the mouth and a key point on another end of the object away from the mouth. For example, the key points near the mouth are defined as p1 and p2 respectively, and the key points away from the mouth are defined as p3 and p4 respectively. It is assumed that the midpoint between p1 and p2 is p5 and the midpoint between p3 and p4 is p6. In this case, the coordinates of p5 and p6 may be used to determine the length of the cigarette.

[0062] In an example, in response to the length of the object interacting with mouth being less than the preset value, it is determined that the image in the first region fails to pass the screening. Then, It may be determined that a cigarette is not involved in the image in the first region.

[0063] Since a large difficulty in the detection of smoking action is how to differentiate a small portion of the cigarette on the image (i.e., when the cigarette substantially exposes only one cross section) and the driver is not in the state of smoking, the feature extracted by the neural network needs to capture very slight detail of the mouth in the image. When the network is required to detect sensitively an object with only exposing a cross section in an image, the false detection rate of the algorithm will be increased. Therefore, in the embodiment of the present disclosure, based on the first key point of the object interacting with mouth, the image in which the portion of the object interacting with mouth is slightly exposed or nothing in the driver's mouth is directly filtered, before sending to the classification network. By testing the trained

network, it is found that in the key point detection algorithm, after the deep network uses the gradient backpropagation algorithm to update the network parameters, it will focus on the edge information of the object interacting with mouth in the image. When some people do not smoke and no strips are around the mouth that will interfere, the prediction of key points will tend to be distributed in an average position in the center of the mouth (even if there is no cigarette at this time). According to the above characteristics, the first key point is used to filter the image that only a small portion of the object interacting with mouth is exposed or that there is nothing on the driver's mouth (that is, the object interacting with mouth only shows a small portion, approximately showing only a cross section of the object, the smoking judgment on the image is insufficient, and it is considered that cigarettes are not involved in the first region).

[0064] In an example, step 240 further includes:

[0065] A serial number is assigned to each of the at least two first key points for distinguishing each first key point.

[0066] By assigning different serial numbers to the first key points, each first key point may be distinguished, and different purposes are achieved by different first key points. For example, the length of the current object may be determined by a first key point closest to the mouth and another first key point farthest from the mouth. The embodiments of the present disclosure may assign serial numbers to the first key points according to an arbitrary non-repetitive order to distinguish different first key points. The embodiments of the present disclosure do not limit the manner of assigning serial numbers, for example, each serial number is assigned to each first key point according to an order of a fork multiplication rule.

[0067] In one or more embodiments, based on the at least two first key points, key

point coordinates corresponding to the at least two first key points are determined in the image in the first region, includes:

[0068] key point coordinates corresponding to the at least two first key points are determined in the image in the first region by using a first neural network.

5 [0069] The first neural network is trained with a first sample image.

[0070] In an example, the first sample image comprises labelled key point coordinates;

[0071] The process of training the first neural network comprises:

[0072] the first sample image is input into the first neural network to obtain predicted key point coordinates corresponding to the at least two first key points;

10 [0073] a first network loss is determined based on the predicted key point coordinates and the labelled key point coordinates, and a parameter of the first neural network is adjusted based on the first network loss.

[0074] In an example, a first key point positioning task, similar to a face key point positioning task, may be regarded as a regression task, so as to obtain a mapping
15 function of the two-dimensional coordinates (x_i, y_i) of the first key point, and the algorithm is described as follows.

[0075] The input of layer 1 of the first neural network is denoted as x_i (i.e., an input image), the output of an intermediate layer of the first neural network is x_n , and each layer of network is equivalent to an nonlinear mapping $F(x)$. Assuming that the first
20 neural network has a total of N layers, then after passing through the nonlinear mapping of the first neural network, the output of the first neural network may be abstracted as the expression of formula 1:

$$\hat{y} = F^N(x) \quad \text{formula 1}$$

[0076] where $\hat{y}$ is a one-dimensional vector output by the first neural network, and
25 each value in the one-dimensional vector represents a key point coordinate finally

output by the network.

[0077] In one or more embodiments, step 230 includes:

[0078] key points of objects that interacts with the mouth are identified for the images in the first region, and the at least two central axis key points on a central axis of the object interacting with mouth, and/or at least two side key points on each of the two sides of the objects that interacts with the mouth.

[0079] In the embodiment of the present disclosure, when the first key point is defined, the central axis key points on the central axis of the object interacting with mouth in the image may be used as the first key points, and/or the side key points on each of the two sides of the object interacting with mouth in the image may be used as the first key points. In order to perform subsequent key point alignment, the key points on each of the two sides is taken as an example. FIG. 3A is a schematic diagram of a first key point obtained via recognition in action recognition method according to an embodiment of the present disclosure. FIG. 3B is a schematic diagram of a first key point obtained via recognition in an action recognition method according to an embodiment of the present disclosure. As shown in FIG. 3A and FIG. 3B, the key points on each of the two sides are defined as a first key points, and in order to identify different first key points and obtain key point coordinates corresponding to different first key points, different serial numbers may be assigned to each of the first key points.

[0080] FIG. 4 is another schematic flowchart of an action recognition method according to an embodiment of the present disclosure. As shown in FIG. 4, the method in this embodiment includes the following steps.

[0081] At step 410, mouth key points of a face are obtained based on the face image.

[0082] At step 420, an image in a first region is determined based on the mouth key points.

[0083] At step 430, at least two second key points on the object interacting with mouth are obtained based on the image in the first region.

[0084] In an example, the at least second key points obtained in this embodiment and the first key points in the above embodiment are both key points on an object interacting with mouth, and the second key points may be the same as or different from the first key points.

[0085] At step 440, the object interacting with mouth is aligned based on the at least two second key points in a way that the object interacting with mouth is oriented to a preset direction, and an image in a second region involving the object interacting with mouth and oriented to the preset direction is obtained.

[0086] The image in the second region involves at least part of the mouth key points and comprises an image of an object interacting with the mouth.

[0087] In the embodiment of the present disclosure, the object interacting with mouth is aligned by obtaining the second key points. so that the object interacting with mouth is oriented to a preset direction; an image in a second region of the object interacting with mouth and is oriented to the preset direction is obtained, the second region and the first region in the above embodiments may have overlapping portions, for example, the second region involves at least part of the mouth key points in the image of the first region and includes the image of the object interacting with mouth. The action recognition method according to the embodiment of the present disclosure may include a plurality of implementations, for example, if only the screening operation is performed on the image in the first region, the first key point of the object interacting with mouth needs to be determined, and the object interacting with mouth is aligned based on the at least two second key points. If the alignment operation is performed on only the object interacting with mouth, the second key point of the object interacting

with mouth needs to be determined, and the alignment operation is performed on the object interacting with mouth based on at least two second key points. If both the screening operation and the alignment operation are performed, the first key points and the second key points of the object interacting with mouth need to be determined, where
5 the first key points may be the same as or different from the second key points, and the second key points and the coordinates thereof may be determined by referring to the first key points and the coordinates thereof, and the operation sequence of the screening operation and the alignment operation is not limited in the embodiment of the present disclosure.

10 **[0088]** In an example, step 440 may include that, corresponding key point coordinates are obtained based on at least two second key points, and an alignment operation is performed based on the obtained key point coordinates corresponding to the second key points. The process of obtaining the key point coordinates based on the second key points may be, similar to the process of obtaining the key point coordinates based on
15 the first key points, for example, obtained by the neural network, the embodiment of the present disclosure does not limit the specific manner including the alignment operation based on the second key points.

[0089] In an example, step 440 may further include that, a serial number for distinguishing each second key point is assigned to each of the at least two second key
20 points. The rule of assigning the serial number may refer to the manner of assigning the serial number to the first key point, which is not described herein.

[0090] At step 450, it is determined whether the person corresponding to the face image is smoking based on the image in the second region.

[0091] Due to the poor rotation invariance of convolutional neural networks, the
25 feature extraction of the neural network at different degrees of rotation of the object has

a certain difference. When a person is smoking, the direction of the cigarette may be in different directions, and if the feature extraction is performed directly on the original captured image, the detection performance of smoking or not may occur to a certain extent. In other words, the neural network needs to adapt to the feature extraction of cigarette at different angles, so as to perform a certain degree of decoupling. In the embodiment of the present disclosure, the alignment operation is performed based on the second key points, so that the objects that interacts with the mouth in each input face image are directed in the same direction, which can reduce the probability of false detection.

5

10 **[0092]** In an example, the alignment operation may include:

[0093] key point coordinates are obtained based on the at least two second key points, and the object interacting with mouth is obtained based on key point coordinates corresponding to the at least two second key points;

[0094] an alignment operation on the object interacting with mouth is performed

15 based on the preset direction by using affine transformation, so that the object interacting with mouth is oriented to the preset direction, and the image in the second region involving the object interacting with mouth and oriented to the preset direction is obtained.

[0095] The affine transformation may include, but is not limited to, at least one of the

20 following: rotation, scaling, translation, flipping, shearing, etc.

[0096] In the embodiment of the present disclosure, pixels in the image of the object interacting with mouth are mapped to a new image aligned by the key points via the affine transformation. The original second key points are aligned with the previously preset key points. In this way, the signal of the object interacting with mouth in the

25 image and the angle information of the object interacting with mouth may be decoupled,

thereby improving the feature extraction performance of the subsequent neural network.

FIG. 5 is a schematic diagram of performing an alignment operation on an object

interacting with mouth in an action recognition method according to another

embodiment of the present disclosure. As shown in FIG. 5, the direction of the object

5 interacting with mouth in the first region image is transformed by performing the affine

transformation using the second key points and the target position, and in this example,

the direction of the object (cigarette) that interacts with the mouth is turned downward.

[0097] The key point alignment is performed by the affine Transformation. The

function of the affine transformation is a linear transformation from two-dimensional

10 coordinates to two-dimensional coordinates, keeping 'straightness' and 'parallelism' of

the two-dimensional graph. The affine transformation may be implemented by a

combination of a series of atomic transformation, where the atomic transformation may

include, but is not limited to, translation, scaling, flipping, rotation, shearing, etc.

[0098] The second coordinate system of the affine transformation is shown in formula

15 (2):

$$[x' \ y' \ 1] = [x \ y \ 1] \cdot \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{12} & a_{22} & 0 \\ x_0 & y_0 & 1 \end{bmatrix} \quad \text{formula (2)}$$

[0099] where, $[x' \ y' \ 1]$ represents the coordinates obtained after the affine

transformation, $[x \ y \ 1]$ represents the key point coordinates of the extracted

cigarette key points, $\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$ represents the rotation matrix, and, x_0 and y_0

20 represent the translation vectors.

[0100] The above expressions encompass several operations of rotation, translation,

scaling, and rotation. Assuming that the key points given by the model are a set of

(x_i, y_i) , the preset target point position (x_i', y_i') (the target point position here may be

preset by human implementation), the affine transformation matrix performs the affine transformation on the source image to the target image, and after capturing, the regular image is obtained.

[0101] In an example, step 130 includes:

5 **[0102]** it is determined whether the person corresponding to the face image is smoking based on the image in the first region by the second neural network.

[0103] The second neural network is trained with the second sample image. The second sample image comprises a sample image of smoking and a sample image of non-smoking, so that the neural network may be trained to distinguish cigarettes from
10 other elongated objects, thereby identifying whether they are smoking or having something else in their mouth.

[0104] In the embodiment of the present disclosure, the obtained key point coordinates are input to the second neural network for classification (for example, a classification convolutional neural network). As an example, the operation process is
15 also performed by the convolutional neural network for feature extraction, and the result of the binary (2-class) classification is finally output, that is, the probability that the image belongs to a smoking or non-smoking image is fitted.

[0105] In an example, the second sample image is associated with a label of whether the person corresponding to the image is smoking;

20 **[0106]** The process of training the second neural network comprises:

[0107] inputting the second sample image into the second neural network, to obtain a prediction of whether the person corresponding to the second sample image is smoking;

[0108] obtaining a second network loss based on the prediction and the label, and adjusting a parameter of the second neural network based on the second network loss.

25 **[0109]** In an example, during the process of training the second neural network, the

network supervision may adopt a softmax loss function, and the mathematical expression is as follows.

[0110] p_i represents the probability that the prediction of the i^{th} second sample image output by the second neural network is the actual correct category (i.e., label), and N represents the total number of samples.

[0111] The loss function may adopt the following formula (3):

$$L_{\text{softmax}} = -\frac{1}{N} \sum_{i=1}^N \log(p_i) \quad \text{formula (3)}$$

[0112] After the network structure and the loss function are defined, training only needs to update the network parameters according to the calculation method of gradient back propagation, to obtain the network parameters of the second neural network after training.

[0113] After the second neural network is trained, the loss function is removed and the fixed network parameter is unchanged, and the pre-processed image is also input to the convolutional neural network to extract features and classifications, so that the classification result given by the classification module may be obtained. Thus, it is determined whether the person in the screen is smoking.

[0114] In one or more embodiments, step 110 includes:

[0115] face key point extraction is performed on a face image to obtain face key points in the face image;

[0116] the mouth key points are obtained based on the face key points.

[0117] In an example, the face key point extraction is performed on the face image by the neural network, and since the smoking action and the interaction mode of the person are mainly performed with a mouth and a hand, the smoking action is basically in the vicinity of the mouth when the smoking action is performed, the valid information region (first region image) may be reduced to the vicinity of the mouth by the face

detection and the face key point positioning technology; In an example, the extracted face key points are edited in serial numbers, and the mouth key points may be obtained by setting the key points of certain serial numbers as mouth key points or by obtaining the mouth key points from the positions of the face key points in the face image, and
5 the first region image is determined based on the mouth key points.

[0118] In some examples, the face image of the embodiment of the present disclosure is obtained by the face detection. The captured image is subjected to the face detection to obtain a face image. The face detection is an underlying infrastructure module for whole smoking action recognition. Since a smoker has a face on the image when he is
10 smoking, the position of the face can be roughly located by the face detection, and the embodiment of the present application does not limit the specific face detection algorithm.

[0119] After the face frame is obtained by the face detection, an image in the face frame (corresponding to the face image in the above embodiment) is captured and the
15 face key points are extracted. In an example, the face key point positioning task may actually be abstracted as a regression task: given an image containing face information, a mapping function of the two-dimensional coordinates (x_i, y_i) of the key points in the image is fitted; for an input image, the detected face position is captured, and the fitting of the network is performed only within a range of one local image, thereby improving
20 the fitting speed. Face key points mainly include key points of five senses. The embodiments of the present disclosure mainly focus on key points of mouth, for example, mouth corner points, lip contour key points, etc.

[0120] In an example, an image in a first region is determined based on the mouth key points includes:

25 **[0121]** a center of the mouth is determined based on the mouth key points;

[0122] the first region is determined by taking the center of the mouth as a center point of the first region and taking a preset length as a side length or a radius.

[0123] In the embodiments of the present disclosure, in order to include the region where cigarettes may appear in the first region, the center of the mouth is determined
5 as the center point of the first region of the image, and a first region of a rectangle or a circle is determined by using a preset length as a radius or a side length. As an example, the preset length may be preset in advance, or determined according to the distance between the center of the mouth and a certain key point in the face. For example, the preset length may be determined based on the distance between one of the mouth key
10 points and the eyebrow key point.

[0124] In an example, the eyebrow key points are obtained based on face key points;

[0125] the first region is determined by taking the center of the mouth as the center point of the first region, and taking a preset length as a side length or a radius, comprising:

15 [0126] the first region is determined by taking the center of the mouth as the center point and taking a vertical distance from the center of the mouth to a center of an eyebrow as a side length or a radius.

[0127] Where the center of the eyebrow is determined based on the at least one eyebrow key point.

20 [0128] For example, after positioning the face key points, the vertical distance d between the center of the mouth and the center of the eyebrow is calculated, then a square region R with the center of the mouth as the center of the square region and $2d$ as the side length is obtained, and the region R is taken as the first region of the embodiment of the present disclosure.

25 [0129] FIG. 6A is a captured original image of the action recognition method

according to the embodiment of the present disclosure. FIG. 6B is a schematic diagram of detecting a face frame in an action recognition method according to an embodiment of the present disclosure. FIG. 6C is a schematic diagram of a first region determined based on key points in the action recognition method according to an embodiment of the present disclosure. In an example, referring to FIGs. 6A, 6B and 6C, the process of obtaining the first region based on the captured original image is implemented.

[0130] A person of ordinary skill in the art may understand that all or portion of the steps of the method embodiments may be implemented by hardware related to program instructions, and the program may be stored in a computer readable storage medium, and when the program is executed, the steps of the method embodiments are performed. The storage medium includes any medium that may store program codes, such as a ROM, a RAM, a magnetic disk, or an optical disk.

[0131] FIG. 7 is a schematic structural diagram of an action recognition apparatus according to an embodiment of the present disclosure. The apparatus in this embodiment may be configured to implement the foregoing method embodiments of the present disclosure. As shown in FIG. 7, the apparatus in this embodiment includes: a mouth key point module 71, a first region determining module 72 and a smoking recognizing module 73.

[0132] The mouth key point module 71 is configured to obtain mouth key points of a face based on a face image.

[0133] The first region determining module 72 is configured to determine an image in a first region based on the mouth key points.

[0134] The image in the first region involves at least part of the mouth key points and comprises an image of an object interacting with mouth.

[0135] The smoking recognizing module 73 is configured to determine whether a

person corresponding to the face image is smoking based on the image in the first region.

[0136] Based on the action recognition apparatus according to the above embodiment of the present disclosure, mouth key points of a face is obtained based on a face image; an image in a first region is determined based on the mouth key points, the image in the first region includes at least part of the mouth key points and an object interacting with mouth; based on the image in the first region, it is determined whether the person corresponding to the face image is smoking, and the first region determined by the mouth key point identifies whether the person is smoking, which reduces the recognition range, and concentrates attention on the mouth and the object interacting with mouth, thereby improving the detection rate, reducing the false detection rate, and improving the accuracy of smoking recognition.

[0137] In one or more embodiments, the apparatus further includes: a first key point module and an image filtering module.

[0138] The first key point module is configured to obtain at least two first key points on the object interacting with mouth based on the image in the first region.

[0139] The image filtering module, configured to screen the image in the first region based on the at least two first key points, to select out the image in the first region in which the object interacting with mouth and having a length greater than or equal to a preset value is involved.

[0140] The smoking recognizing module 73 is configured to determine whether the person corresponding to the face image is smoking based on the image in the first region, in response to that the image in the first region passes the screening.

[0141] In an example, the image filtering module is configured to, determine key point coordinates corresponding to the at least two first key points in the image in the first region based on the at least two first key points, and screen the image in the first

region based on the key point coordinates.

[0142] In an example, the image filtering module is configured to, when screening the image in the first region based on the key point coordinates, determine a length of the object interacting with mouth in the image in the first region based on the key point coordinates; and in response to the length of the object interacting with mouth being greater than or equal to the preset value, determine that the image in the first region passes the screening.

[0143] In an example, the image filtering module is further configured to, when screening the image in the first region based on the key point coordinates, in response to the length of the object interacting with mouth being less than the preset value, determine that the image in the first region fails to pass the screening, and determine that the image in the first region does not involve a cigarette.

[0144] In an example, the image filtering module is further configured to assign a serial number to each of the at least two first key points to distinguish the at least two first key points.

[0145] In an example, the image filtering module is configured to, determine, by a first neural network, the key point coordinates corresponding to the at least two first key points in the image in the first region, wherein the first neural network is trained with a first sample image, when determining the key point coordinates corresponding to the at least two first key points in the image in the first region.

[0146] In an example, the first sample image includes labelled key point coordinates; training the first neural network includes:

[0147] inputting the first sample image into the first neural network to obtain predicted key point coordinates corresponding to the at least two first key points;

[0148] determining a first network loss based on the predicted key point coordinates

and the labelled key point coordinates, and adjusting a parameter of the first neural network based on the first network loss.

[0149] In an example, the first key point module is configured to perform a key point recognition for the object interacting with mouth on the image in the first region, and
5 obtain at least two central axis key points on a central axis of the object interacting with mouth, and/or at least two side key points on each of the two sides of the object interacting with mouth.

[0150] In one or more embodiments, the apparatus according to the embodiments of the present disclosure further includes a second key point module, and an image
10 aligning module.

[0151] The second key point module is configured to obtain at least two second key points on the object interacting with mouth based on the image in the first region.

[0152] The image aligning module is configured to align the object interacting with mouth based on the at least two second key points in a way that the object interacting
15 with mouth is oriented to a preset direction, obtain an image in a second region of the object interacting with mouth and oriented to the preset direction, wherein the image in the second region includes at least part of the mouth key points and the object interacting with mouth.

[0153] The smoking recognizing module 73 is configured to determine whether the
20 person corresponding to the face image is smoking based on the image in the second region.

[0154] In one or more embodiments, the smoking recognizing module 73 is configured to determine, by a second neural network, whether the person corresponding to the face image is smoking based on the image in the first region, wherein the second
25 neural network is trained with a second sample image.

[0155] In an example, the second sample image is associated with a label of whether the person corresponding to the image is smoking; training the second neural network includes:

[0156] inputting the second sample image into the second neural network, to obtain a
5 prediction of whether the person corresponding to the second sample image is smoking;

[0157] obtaining a second network loss based on the prediction and the label, and adjusting a parameter of the second neural network based on the second network loss.

[0158] In one or more embodiments, the mouth key point module 71 is configured to perform face key point extraction on the face image to obtain a plurality of face key
10 points in the face image; and obtain the mouth key points based on the plurality of face key points.

[0159] In an example, the first region determining module 72 is configured to determine a center of the mouth based on the mouth key points; and determine the first region by taking the center of the mouth as a center point of the first region and taking
15 a preset length as a side length or a radius.

[0160] In an example, the apparatus according to the embodiment of the present disclosure further includes an eyebrow key point module.

[0161] The eyebrow key point module is configured to obtain at least one eyebrow key point based on the face key points.

20 [0162] The first region determining module 72 is configured to determine the first region by taking the center of the mouth as the center point of the first region, and taking a vertical distance from the center of the mouth to an eyebrow as a side length or a radius, wherein the eyebrow is determined based on the eyebrow key point.

[0163] The working process, the setting mode, and the corresponding technical effects
25 of any embodiment of the action recognition apparatus according to the embodiments

of the present disclosure may be referred to the specific description of the above corresponding method embodiments of the present disclosure, which will not be described herein again.

[0164] According to another aspect of the embodiments of the present disclosure, an electronic device is provided, which includes a processor. The processor includes the action recognition apparatus according to any of the above embodiments.

[0165] According to another aspect of the embodiments of the present disclosure, an electronic device is provided, including: a memory configured to store executable instructions; and

[0166] a processor for communicating with the memory to execute executable instructions to complete operations in the action recognition method according to any of the above embodiments.

[0167] According to another aspect of the embodiments of the present disclosure, a computer readable storage medium is provided, configured to store computer readable instructions, wherein the instructions are executed to perform the operations in the action recognition method according to any of the above embodiments.

[0168] According to another aspect of the embodiments of the present disclosure, a computer program product is provided, including computer readable codes, wherein the computer readable codes are running on a device to cause a processor in the device to execute instructions for implementing the action recognition method according to any of the above embodiments.

[0169] The embodiment of the present disclosure further provides an electronic device, for example, may be a mobile terminal, a personal computer (PC), a tablet computer, a server, etc. Referring now to FIG. 8, which shows a schematic structural diagram of an electronic device 800 suitable for implementing a terminal device or a

server according to an embodiment of the present disclosure, as shown in FIG. 8. The electronic device 800 includes one or more processors, a communication unit, and the like. The one or more processors may be, for example, one or more central processing units (CPUs) 801, and/or one or more image processors (acceleration units) 813, etc.

5 The processor may perform various appropriate actions and processes according to executable instructions stored in the read-only memory (ROM) 802 or executable instructions loaded from the storage section 808 into the random access memory (RAM) 803. The communication unit 812 may include, but is not limited to, a network card, and the network card may include, but is not limited to, an IB (Infiniband) network card.

10 **[0170]** The processor may communicate with the read-only memory 802 and/or the random access memory 803 to execute executable instructions, connect with the communication unit 812 via the bus 804, and communicate with other target devices via the communication unit 812, thereby completing operations corresponding to any method according to the embodiments of the present disclosure, for example, obtaining

15 a plurality of mouth key points of a face based on a face image; determining an image in a first region based on the plurality of mouth key points, the image in the first region involving at least part of the plurality of mouth key points and an object interacting with mouth; determining whether a person corresponding to the face image is smoking based on the image in the first region.

20 **[0171]** The RAM 803 may also store various programs and data required for device operation. The CPU 801, the ROM 802, and the RAM 803 are connected to each other via a bus 804. In the case where the RAM 803 is present, the ROM 802 is an optional module. The RAM 803 stores an executable instruction, or writes an executable instruction into the ROM 802 at runtime, and the executable instruction causes the

25 central processing unit 801 to execute an operation corresponding to the foregoing

communication method. The input/output (I/O) interface 805 is also connected to the bus 804. The communication unit 812 may be integrally arranged, or may be arranged to have a plurality of sub-modules (for example, a plurality of IB network cards) and be connected to a bus link.

5 **[0172]** The following components are connected to the I/O interface 805: an input section 806 including a keyboard, a mouse, etc. An output section 807 such as a cathode ray tube (CRT), a liquid crystal display (LCD), and the like, and a speaker. a storage section 808 including a hard disk or the like, and a communication section 809 including a network interface card such as a LAN card, a modem or the like. The
10 communication section 809 performs communication processing via a network such as the Internet. Driver 810 is also connected to I/O interface 805 as needed. A detachable medium 811, such as a magnetic disk, an optical disk, a magneto-optical disk, a semiconductor memory, or the like, is mounted on the driver 810 as needed, so that a computer program read out therefrom is mounted in the storage section 808 as needed.

15 **[0173]** The architecture shown in FIG. 8 is merely an optional implementation, and during specific practice, the number and type of the components shown in FIG. 8 may be selected, deleted, added or replaced according to actual needs. Implementations such as separation setting or integration setting may also be adopted on different functional component arrangements, for example, the acceleration unit 813 and the CPU 801 may
20 be separately arranged or the acceleration unit 813 may be integrated on the CPU 801, the communication unit may be separately arranged, or may be integrated on the CPU 801 or the acceleration unit 813, etc. These alternative embodiments all belong to the scope of protection disclosed herein.

25 **[0174]** In particular, according to embodiments of the present disclosure, the processes described above with reference to the flowcharts may be implemented as

computer software programs. For example, an embodiment of the present disclosure comprises a computer program product comprising a computer program tangibly contained on a machine readable medium, the computer program comprising program codes for executing the method shown in the flowchart, and the program codes may
5 comprise instructions corresponding to the step of executing the method provided in the embodiment of the present disclosure, for example, obtaining a mouth key point of a face based on a face image; determining an image in a first region based on a mouth key point, the image in the first region including at least part of a mouth key point and an image of an object interacting with mouth; determining whether a person
10 corresponding to the face image is smoking based on the image in the first region. In such embodiments, the computer program may be downloaded and installed from the network through the communication section 809 and/or installed from the removable medium 811. When the computer program is executed by the central processing unit (CPU) 801, the operation of the described function defined in the method of the present
15 disclosure is performed.

[0175] Different examples in the present disclosure are described in a progressive manner. Each example focuses on the differences from other examples with those same or similar parts among the examples to be referred to each other. Particularly, since the data processing device examples are basically similar to the method examples, the
20 device examples are briefly described with relevant parts referred to the descriptions of the method examples.

[0176] The methods and apparatuses of the present disclosure may be implemented in a number of ways. For example, the methods and apparatuses of the present disclosure may be implemented by software, hardware, firmware, or any combination
25 of software, hardware, firmware. The above-mentioned order of the steps for the

method is merely for illustration, and the steps of the method of the present disclosure are not limited to the order specifically described above, unless specifically stated otherwise. Furthermore, in some embodiments, the present disclosure may also be embodied as programs recorded in a recording medium, including machine readable
5 instructions for implementing the method according to the present disclosure. Accordingly, the present disclosure also covers a recording medium storing a program for executing the method according to the present disclosure.

[0177] The description of the present disclosure is given for purposes of example and description and is not missing or limiting the present disclosure to the disclosed forms.
10 Many modifications and variations will be apparent to those skilled in the art. The embodiments are chosen and described in order to better illustrate the principles and practical applications of the present disclosure, and to enable those skilled in the art to understand the present disclosure in order to design various embodiments with various modifications suitable for specific purpose.

15

CLAIMS

1. An action recognition method, comprising:
 - obtaining mouth key points of a face based on a face image;
 - 5 determining, based on the mouth key points, an image in a first region involves at least part of the mouth key points and comprises an image of an object interacting with mouth; and
 - determining whether a person corresponding to the face image is smoking based on the image in the first region.
- 10 2. The method according to claim 1, wherein before determining whether the person corresponding to the face image is smoking based on the image in the first region, the method further comprises:
 - obtaining at least two first key points of the object interacting with mouth based on
 - 15 the image in the first region; and
 - screening the image in the first region based on the at least two first key points, to select out the image in the first region in which the object interacting with mouth and having a length greater than or equal to a preset value is involved,
 - determining whether the person corresponding to the face image is smoking based
 - 20 on the image in the first region comprises:
 - in response to that the image in the first region passes the screening,
 - determining whether the person corresponding to the face image is smoking based on the image in the first region.
- 25 3. The method according to claim 2, wherein screening the image in the first region

based on the at least two first key points comprises:

determining key point coordinates corresponding to the at least two first key points in the image in the first region; and

screening the image in the first region based on the key point coordinates
5 corresponding to the at least two first key points.

4. The method according to claim 3, wherein screening the image in the first region based on the key point coordinates corresponding to the at least two first key points comprises:

10 determining, based on the key point coordinates corresponding to the at least two first key points, a length of the object interacting with mouth which is involved in the image in the first region; and

in response to that the length of the object interacting mouth is greater than or equal to the preset value, determining that the image in the first region passes the screening.

15

5. The method according to claim 4, further comprising:

in response to that the length of the object interacting with mouth is less than the preset value, determining that the image in the first region fails to pass the screening; and

20 determining that the image in the first region does not involve a cigarette.

6. The method according to any of claims 3 to 5, wherein before determining the key point coordinates corresponding to the at least two first key points in the image in the first region, the method further comprises:

25 assigning a serial number to each of the at least two first key points to distinguish

the at least two first key points.

7. The method according to any of claims 3 to 6, wherein determining the key point coordinates corresponding to the at least two first key points in the image in the first region comprises:

determining, by a first neural network, the key point coordinates corresponding to the at least two first key points in the image in the first region, wherein the first neural network is trained with a first sample image.

8. The method according to claim 7, wherein the first sample image comprises labelled key point coordinates, and training the first neural network comprises:

inputting the first sample image into the first neural network to obtain predicted key point coordinates corresponding to the at least two first key points;

determining a first network loss based on the predicted key point coordinates and the labelled key point coordinates; and

adjusting a parameter of the first neural network based on the first network loss.

9. The method according to any of claims 2 to 8, wherein obtaining the at least two first key points of the object interacting with mouth based on the image in the first region comprises:

performing a key point recognition for the object interacting with mouth on the image in the first region to obtain

at least two central axis key points on a central axis of the object interacting with mouth, and/or

at least two side key points on each of two sides of the object interacting with

mouth.

10. The method according to any of claims 1 to 9, wherein before determining whether the person corresponding to the face image is smoking based on the image in the first
5 region, the method further comprises:

obtaining at least two second key points of the object interacting with mouth based on the image in the first region,

aligning, based on the at least two second key points, the object interacting with mouth in a way that the object interacting with mouth is oriented to a preset direction,
10 and

obtaining an image in a second region involving the object interacting with mouth and oriented to the preset direction, wherein the image in the second region involves at least part of the mouth key points and comprises an image of the object interacting with mouth;

15 determining whether the person corresponding to the face image is smoking based on the image in the first region comprises:

determining whether the person corresponding to the face image is smoking based on the image in the second region.

20 11. The method according to any of claims 1 to 10, wherein determining whether the person corresponding to the face image is smoking based on the image in the first region comprises:

determining, by a second neural network, whether the person corresponding to the face image is smoking based on the image in the first region, wherein the second neural
25 network is trained with a second sample image.

12. The method according to claim 11, wherein the second sample image is associated with a label of whether the person corresponding to the second sample image is smoking, and training the second neural network comprises:

- 5 inputting the second sample image into the second neural network to obtain a prediction of whether a person corresponding to the second sample image is smoking;
 obtaining a second network loss based on the prediction and the label, and
 adjusting a parameter of the second neural network based on the second network loss.

10

13. The method according to any of claims 1 to 12, wherein obtaining the mouth key points of the face based on the face image comprises:

- performing a face key point extraction on the face image to obtain face key points in the face image;
- 15 obtaining the mouth key points based on the face key points.

14. The method according to claim 13, wherein determining the image in the first region based on the mouth key points comprises:

- determining a center of a mouth involved in the face image based on the mouth key
- 20 points;
- determining the first region by taking the center of the mouth as a center point of the first region and taking a preset length as a side length or a radius.

15. The method according to claim 14, wherein before determining the image in the
- 25 first region based on the mouth key points, the method further comprises:

obtaining at least one eyebrow key point based on the face key points;

determining the first region by taking the center of the mouth as the center point of the first region and taking the preset length as the side length or the radius comprises:

5 determining the first region by taking the center of the mouth as the center point of the first region, and taking a vertical distance from the center of the mouth to a center of an eyebrow as the side length or the radius, wherein the center of the eyebrow is determined based on the at least one eyebrow key point.

16. An action recognition apparatus, comprising:

10 a mouth key point module, configured to obtain mouth key points of a face based on a face image;

a first region determining module, configured to determine, based on the mouth key points, an image in a first region involves at least part of the mouth key points and comprises an image of an object interacting with mouth; and

15 a smoking recognizing module, configured to determine whether a person corresponding to the face image is smoking based on the image in the first region.

17. The apparatus according to claim 16, further comprising:

20 a first key point module, configured to obtain at least two first key points on the object interacting with mouth based on the image in the first region; and

an image filtering module, configured to screen the image in the first region based on the at least two first key points, to select out the image in the first region in which the object interacting with mouth and having a length greater than or equal to a preset value is involved,

25 the smoking recognizing module is configured to, in response to the image in the

first region passing the screening, determine whether the person corresponding to the face image is smoking based on the image in the first region.

18. The apparatus according to claim 17, wherein the image filtering module is
5 configured to:

determine, based on the at least two first key points, key point coordinates corresponding to the at least two first key points in the image in the first region; and
screen the image in the first region based on the key point coordinates corresponding to the at least two first key points.

10

19. The apparatus according to claim 18, wherein the image filtering module is configured to:

when screening the image in the first region based on the key point coordinates corresponding to the at least two first key points,

15 determine, based on the key point coordinates corresponding to the at least two first key points, a length of the object interacting with mouth in the image in the first region based on the key point coordinates; and

in response to that the length of the object interacting with mouth is greater than or equal to the preset value, determine that the image in the first region passes the
20 screening.

20. The apparatus according to claim 19, wherein the image filtering module is further configured to:

when screening the image in the first region based on the key point coordinates,
25 in response to that the length of the object interacting with mouth is less than the

preset value, determine that the image in the first region fails to pass the screening, and determine that the image in the first region does not involve a cigarette.

21. The apparatus according to any of claims 18 to 20, wherein the image filtering
5 module is further configured to:

assign a serial number to each of the at least two first key points to distinguish the
at least two first key points.

22. The apparatus according to any of claims 18 to 21, wherein the image filtering
10 module is configured to:

when determining the key point coordinates corresponding to the at least two first
key points in the image in the first region,

determine, by a first neural network, the key point coordinates corresponding to
the at least two first key points in the image in the first region, wherein the first neural
15 network is trained with a first sample image.

23. The apparatus according to claim 22, wherein

the first sample image comprises labelled key point coordinates; and

training the first neural network comprises:

20 inputting the first sample image into the first neural network to obtain
predicted key point coordinates corresponding to the at least two first key points,
and

determining a first network loss based on the predicted key point coordinates
and the labelled key point coordinates, and adjusting a parameter of the first neural
25 network based on the first network loss.

24. The apparatus according to any of claims 17 to 23, wherein the first key point module is configured to:
- perform a key point recognition for the object interacting with mouth on the image
 - in the first region, to obtain
 - at least two central axis key points on a central axis of the object interacting with mouth, and/or at least two side key points on each of the two sides of the object interacting with mouth.
25. The apparatus according to any of claims 16 to 24, further comprising:
- a second key point module, configured to obtain at least two second key points of the object interacting with mouth based on the image in the first region;
 - an image aligning module, configured to
 - align the object interacting with mouth based on the at least two second key points, such that the object interacting with mouth is oriented to a preset direction,
 - and
 - obtain an image in a second region involving the object interacting with mouth and oriented to the preset direction, wherein the image in the second region involves at least part of the mouth key points and comprises an image of the object interacting with mouth; and
 - the smoking recognizing module is configured to determine whether the person corresponding to the face image is smoking based on the image in the second region.
26. The apparatus according to any of claims 16 to 25, wherein the smoking recognizing module is configured to:

determine, by a second neural network, whether the person corresponding to the face image is smoking based on the image in the first region, wherein the second neural network is trained with a second sample image.

5 27. The apparatus according to claim 26, wherein

the second sample image is associated with a label of whether the person corresponding to the second sample image is smoking; and
training the second neural network comprises:

inputting the second sample image into the second neural network, to obtain a
10 prediction of whether a person corresponding to the second sample image is smoking;

obtaining a second network loss based on the prediction and the label, and
adjusting a parameter of the second neural network based on the second network loss.

15

28. The apparatus according to any of claims 16 to 27, wherein the mouth key point module is configured to

perform a face key point extraction on the face image to obtain face key points in the face image, and

20 obtain the mouth key points based on the face key points.

29. The apparatus according to claim 28, wherein the first region determining module is configured to:

determine a center of mouth involved in the face image based on the mouth key
25 points;

determine the first region by taking the center of the mouth as a center point of the first region and taking a preset length as a side length or a radius.

30. The apparatus according to claim 29, further comprising:

5 an eyebrow key point module, configured to obtain at least one eyebrow key point based on the face key points; and

 the first region determining module is configured to determine the first region by taking the center of the mouth as the center point of the first region, and taking a vertical distance from the center of the mouth to an eyebrow as the side length or the radius,

10 wherein the eyebrow is determined based on the at least one eyebrow key point.

31. An electronic device, comprising: a processor, wherein the processor comprises an action recognition apparatus according to any of claims 16 to 30.

15 32. An electronic device, comprising:

 a memory configured to store executable instructions; and

 a processor configured to communicate with the memory to execute the executable instructions to perform operations in an action recognition method according to any of claims 1 to 15.

20

33. A computer readable storage medium, configured to store computer readable instructions, wherein the instructions are executed to perform the action recognition method according to any of claims 1 to 15.

25 34. A computer program product comprising computer readable codes, wherein the

computer readable codes are running on a device to cause a processor in the device to execute instructions for implementing an action recognition method according to any of claims 1 to 15.